CLAIMS

What is claimed is:

1. A method of transparent content addressable data storage and compression for a file system comprising:

providing a data structure that associates file identifiers and retrieval keys for memory blocks for storing file contents;

storing in the data structure one or more file identifiers;

providing a chunk of data comprising a quantity of input data of a file;

retrieving a memory block from computer memory;

searching for a segment of the chunk that matches the memory block; and

if a matching segment is found:

discarding the matching segment, providing a retrieval key for the memory block as a retrieval key for the matching segment, and storing in the data structure the retrieval key for the matching segment in association with a file identifier; and

identifying an unmatched portion of the chunk that does not match the memory block, storing the unmatched portion, providing a retrieval key for the unmatched portion, and storing in the data structure the retrieval key for the unmatched portion in association with the file identifier.

2. The method of claim 1 further comprising iteratively storing retrieval keys for

each file until each file identifier is associated with only one retrieval key.

3. The method of claim 8 further comprising iteratively storing all file identifiers and associated retrieval keys until an entire file system is represented by a single retrieval key.

4. The method of claim 1 wherein searching for a segment of the chunk that matches the memory block comprises searching at a repeating memory interval through a search section of the chunk for a segment of the chunk that matches the memory block.

5. The method of claim 4 wherein searching at a repeating memory interval through a search section of the chunk for a segment of the chunk that matches the memory block comprises:

calculating a weak checksum for the memory block;

calculating weak checksums for segments of the search section of the chunk;

comparing the weak checksums for the segments with the checksum for the memory block; and

if a segment is found with a weak checksum equal to the weak checksum of the memory block:

calculating a strong checksum for the memory block;

calculating a strong checksum for the segment with the matching weak checksum;

comparing the strong checksum of the memory block and the strong checksum for the segment with the equal weak checksum; and

determining that the search has found a segment having contents that match the contents of the memory block if the strong checksum of the memory block and the strong checksum for the segment with the matching weak checksum are equal.

6. The method of claim 1 wherein storing the unmatched portion of the chunk comprises storing the unmatched portion of the chunk as a new memory block having a memory block size equal to the size of the unmatched portion of the chunk.

7. The method of claim 1 wherein searching for a segment of the chunk that matches the memory block fails to find a matching segment, the method further comprising repeatedly carrying out the following steps for all memory blocks in computer memory until a matching segment is found:

retrieving a next memory block from computer memory; and

searching for a segment of the chunk that matches the next memory block.

8. The method of claim 7 wherein no matching segment is found in any memory block in computer memory, the method further comprising:

storing a search section of the chunk;

providing a retrieval key for the search section of the chunk; and

storing the retrieval key for the search section in association with a file identifier.

9. The method of claim 1 further comprising reading file contents from computer memory for a file comprising an identifier and one or more associated retrieval keys, including:

identifying memory blocks in dependence upon the associated retrieval keys; and

retrieving from memory the identified memory blocks.

10. A system of transparent content addressable data storage and compression for a file system comprising:

means for providing a data structure that associates file identifiers and retrieval keys for memory blocks for storing file contents;

means for storing in the data structure one or more file identifiers;

means for providing a chunk of data comprising a quantity of input data of a file;

means for retrieving a memory block from computer memory;

means for searching for a segment of the chunk that matches the memory block;

means for discarding a matching segment, means for providing a retrieval key for the memory block as a retrieval key for the matching segment, and means for storing in the data structure the retrieval key for the matching segment in association with a file identifier; and

means for identifying an unmatched portion of the chunk that does not match the memory block, means for storing the unmatched portion, means for providing a retrieval key for the unmatched portion, and means for storing in the data structure the retrieval key for the unmatched portion in association with the file identifier.

11. The system of claim 10 further comprising means for iteratively storing retrieval keys for each file until each file identifier is associated with only one retrieval key.

12. The system of claim 11 further comprising means for iteratively storing all file identifiers and associated retrieval keys until an entire file system is represented by a single retrieval key.

13. The system of claim 10 wherein means for searching for a segment of the chunk that matches the memory block comprises means for searching at a repeating memory interval through a search section of the chunk for a segment of the chunk that matches the memory block.

14. The system of claim 13 wherein means for searching at a repeating memory interval through a search section of the chunk for a segment of the chunk that matches the memory block comprises:

means for calculating a weak checksum for the memory block;

means for calculating weak checksums for segments of the search section of the chunk;

means for comparing the weak checksums for the segments with the checksum for the memory block;

means for calculating a strong checksum for the memory block;

means for calculating a strong checksum for the segment with a matching weak checksum;

means for comparing the strong checksum of the memory block and the strong checksum for the segment with the equal weak checksum; and

means for determining that the search has found a segment having contents that match the contents of the memory block if the strong checksum of the memory block and the strong checksum for the segment with the matching weak checksum are equal.

15. The system of claim 10 wherein means for storing the unmatched portion of the chunk comprises means for storing the unmatched portion of the chunk as a new memory block having a memory block size equal to the size of the unmatched portion of the chunk.

16. The system of claim 10 further comprising:

means for retrieving a next memory block from computer memory; and

means for searching for a segment of the chunk that matches the next memory block.

17. The system of claim 16 further comprising:

means for storing a search section of the chunk;

means for providing a retrieval key for the search section of the chunk; and

means for storing the retrieval key for the search section in association with a file identifier.

18. The system of claim 10 further comprising means for reading file contents from computer memory for a file comprising an identifier and one or more associated retrieval keys, including:

means for identifying memory blocks in dependence upon the associated retrieval keys; and means for retrieving from memory the identified memory blocks.

19. A computer program product of transparent content addressable data storage and compression for a file computer program product comprising:

a recording medium;

means, recorded on the recording medium, for providing a data structure that associates file identifiers and retrieval keys for memory blocks for storing file contents;

means, recorded on the recording medium, for storing in the data structure one or more file identifiers;

means, recorded on the recording medium, for providing a chunk of data comprising a quantity of input data of a file;

means, recorded on the recording medium, for retrieving a memory block from computer memory;

means, recorded on the recording medium, for searching for a segment of the chunk that matches the memory block;

means, recorded on the recording medium, for discarding a matching segment, means, recorded on the recording medium, for providing a retrieval key for the memory block as a retrieval key for the matching segment, and means, recorded on the recording medium, for storing in the data structure the retrieval key for the matching segment in association with a file identifier; and

means, recorded on the recording medium, for identifying an unmatched portion of the chunk that does not match the memory block, means, recorded on the recording medium, for storing the unmatched portion, means, recorded on the recording medium, for providing a retrieval key for the unmatched portion, and means, recorded on the recording medium, for storing in the data structure the retrieval key for the unmatched portion in association with the file identifier.

20. The computer program product of claim 19 further comprising means, recorded on the recording medium, for iteratively storing retrieval keys for each file until each file identifier is associated with only one retrieval key.

21. The computer program product of claim 20 further comprising means, recorded on the recording medium, for iteratively storing all file identifiers and associated retrieval keys until an entire file computer program product is represented by a single retrieval key.

22. The computer program product of claim 19 wherein means, recorded on the recording medium, for searching for a segment of the chunk that matches the memory block comprises means, recorded on the recording medium, for searching at a repeating memory interval through a search section of the chunk for a segment of the chunk that matches the memory block.

23. The computer program product of claim 22 wherein means, recorded on the recording medium, for searching at a repeating memory interval through a search section of the chunk for a segment of the chunk that matches the memory block comprises:

means, recorded on the recording medium, for calculating a weak checksum for the memory block;

means, recorded on the recording medium, for calculating weak checksums for segments of the search section of the chunk;

means, recorded on the recording medium, for comparing the weak checksums for the segments with the checksum for the memory block;

means, recorded on the recording medium, for calculating a strong checksum for the memory block;

means, recorded on the recording medium, for calculating a strong checksum for the segment with a matching weak checksum;

means, recorded on the recording medium, for comparing the strong checksum of the memory block and the strong checksum for the segment with the equal weak checksum; and

means, recorded on the recording medium, for determining that the search has found a segment having contents that match the contents of the memory block if the strong checksum of the memory block and the strong checksum for the segment with the matching weak checksum are equal.

24. The computer program product of claim 19 wherein means, recorded on the recording medium, for storing the unmatched portion of the chunk comprises means, recorded on the recording medium, for storing the unmatched portion of the chunk as a new memory block having a memory block size equal to the size of the unmatched portion of the chunk.

25. The computer program product of claim 19 further comprising:

means, recorded on the recording medium, for retrieving a next memory block from computer memory; and

means, recorded on the recording medium, for searching for a segment of the chunk that matches the next memory block.

26. The computer program product of claim 25 further comprising:

means, recorded on the recording medium, for storing a search section of the chunk;

means, recorded on the recording medium, for providing a retrieval key for the search section of the chunk; and

means, recorded on the recording medium, for storing the retrieval key for the search section in association with a file identifier.

27. The computer program product of claim 19 further comprising means, recorded on the recording medium, for reading file contents from computer memory for a file comprising an identifier and one or more associated retrieval keys, including:

means, recorded on the recording medium, for identifying memory blocks in dependence upon the associated retrieval keys; and

means, recorded on the recording medium, for retrieving from memory the identified memory blocks.